The support vector classifier (2)

Manel Martínez-Ramón

ECE, UNM

The KKT conditions

- The KKT conditions are

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n \tag{1}$$

$$C - \alpha_n - \mu_n = 0 \tag{2}$$

$$\sum_{n=1}^{N} \alpha_n y_n = 0 \tag{3}$$

$$\mu_n \xi_n = 0 \tag{4}$$

$$\alpha_n \left(y_n \left(\mathbf{w}^\top \mathbf{x}_n + b \right) - 1 + \xi_n \right) = 0 \tag{5}$$

$$\alpha_n \geq 0,\ \ \mu_n \geq 0,\ \ \xi_n \geq 0 \tag{6}$$

The solution

- From (2) and (4)

$$
\begin{aligned}
C - \alpha_n - \mu_n &= 0 \\
\mu_n \xi_n &= 0
\end{aligned}
$$

  we see that if $\xi_n > 0$ (sample inside the margin or misclasified), then $\alpha_n = C$.
- With (5), we see that if the sample is on the margin, $0 < \alpha_n < C$
- If the sample is well classified and outside the margin, then $\xi_n = 0$, and (5) determines that $\alpha_n = 0$.

The solution

- The estimator $y_k = \mathbf{w}^\top \mathbf{x}_k + b$ can be rewritten by virtue of (1) as

$$y_k = \sum_{n=1}^{N} y_n \alpha_n \mathbf{x}_n^\top \mathbf{x}_k + b$$

or, in matrix notation

$$y_k = \boldsymbol{\alpha}^\top \mathbf{Y} \mathbf{X}^\top \mathbf{x}_k + b$$

- Finally, if we plug (1) in the Lagrangian of slide 6, we have:

$$\text{(A):}\frac{1}{2}||\mathbf{w}||^2 = \sum_{n=1}^{N}\sum_{n'=1}^{N} y_{n'}\alpha_{n'}\mathbf{x_{n'}}^\top \mathbf{x_n}\alpha_n y_n$$

$$\text{(B):}\ -\sum_{n=1}^{N}\alpha_n\left(y_n\left(\mathbf{w}^\top\mathbf{x}_n+b\right)-1+\xi_n\right)=$$

$$-\sum_{n=1}^{N}\alpha_n\left(y_n\left(\sum_{n'=1}^{N}\alpha_{n'}\mathbf{x}_{n'}^\top\mathbf{x}_n+b\right)-1+\xi_n\right)=$$

$$-\sum_{n=1}^{N}\sum_{n'=1}^{N}\alpha_{n'}y_{n'}\mathbf{x}_{n'}^\top\mathbf{x}_n\alpha_n y_n-\sum_{n=1}^{N}\alpha_n y_n b+\sum_{n=1}^{N}\alpha_n-\sum_{n=0}^{N}\alpha_n\xi_n$$

$$\text{(C):}\ -\sum_{n=1}^{N}\mu_n\xi_n \quad \text{(D):}\ C\sum_{n=1}^{N}\xi_n$$

- Term (C) can be removed by virtue of KKT (4) (see page 9). Then, terms (A), (B) and (D) add to

$$L_d(\alpha_n, \xi_n) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{n'=1}^{N}\alpha_{n'}y_{n'}\mathbf{x}_{n'}^\top\mathbf{x}_n\alpha_n y_n - \sum_{n=1}^{N}\alpha_n y_n b + \sum_{n=1}^{N}\alpha_n - \sum_{n=0}^{N}\alpha_n\xi_n + C\sum_{n=1}^{N}\xi_n$$

Term $\sum_{n=1}^{N}\alpha_n y_n b$ is nulled by KKT number (3). Thus:

$$L_d(\alpha_n, \xi_n) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{n'=1}^{N}\alpha_{n'}y_{n'}\mathbf{x}_{n'}^\top\mathbf{x}_n\alpha_n y_n + \sum_{n=1}^{N}\alpha_n - \sum_{n=0}^{N}\alpha_n\xi_n + C\sum_{n=1}^{N}\xi_n$$

The solution

- We can say that $-\sum_{n=0}^{N}\alpha_n\xi_n + C\sum_{n=1}^{N}\xi_n = 0$. Indeed, if $0 \leq \alpha_n < C$ then $\xi_n = 0$ as explained in slide C, so the sum can be rewriten as

$$-\sum_{\xi_n>0}\alpha_n\xi_n + C\sum_{xi_n>0}^{N}\xi_n$$

but since when $\xi_n > 0$ the corresponding Lagrange multiplier is $\alpha_n = C$, both terms are equal.

- Finally, we have the result

$$L_d = -\frac{1}{2}\sum_{n=1}^{N}\sum_{n'=1}^{N} y_n \alpha_n \mathbf{x_n}^\top \mathbf{x_{n'}} \alpha_{n'} y_{n'} + \sum_{n=1}^{N} \alpha_n$$

  which is, in matrix notation

$$L_d = -\frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{Y}\mathbf{X}^\top \mathbf{X}\mathbf{Y}\boldsymbol{\alpha} + \boldsymbol{\alpha}^\top \mathbf{1}$$

  with the constraint $\boldsymbol{\alpha} \geq \mathbf{0}$.

- The dual functional must be optimized wrt the dual variables using quadratic programming, implemented in many packages, including Matlab or LIB-SVM.

The solution

- The product $\mathbf{X}^\top\mathbf{X}$ is a Gram matrix of dot products, usually notated as $\mathbf{K}$, where:

$$K_{i,j} = \mathbf{x}_i^\top \mathbf{x}_j$$

- The dual is then written as

$$L_d = -\frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{YKY}\boldsymbol{\alpha}^\top + \boldsymbol{\alpha}\mathbf{1}$$

- Matrix $\mathbf{YKY}$ is positive definite, this is, all its eigenvalues are positive, hence

$$\boldsymbol{\alpha}^\top \mathbf{YKY}\boldsymbol{\alpha} > 0$$

which ensures existence and uniqueness of solutions.

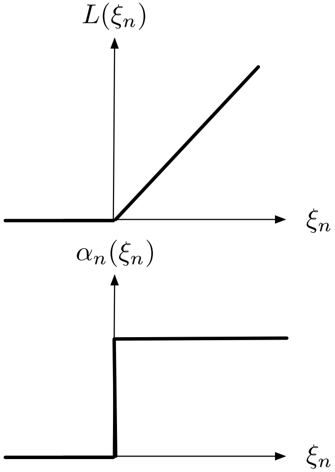
Cost function

The cost function or risk function following the SLT community is

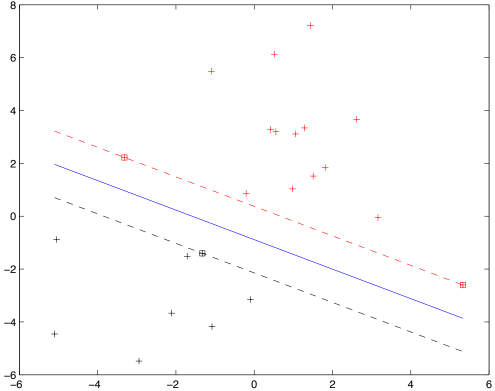
$$L(\xi_n) = \begin{cases} C\xi_n & \xi_n \geq 0 \\ 0 & \xi_n \leq 0 \end{cases}$$

The Lagrange multipliers in the optimal point are derivative of the cost function:

$$\alpha_n(\xi_n) = \begin{cases} C & \xi_n > 0 \\ 0 & \xi_n < 0 \end{cases}$$

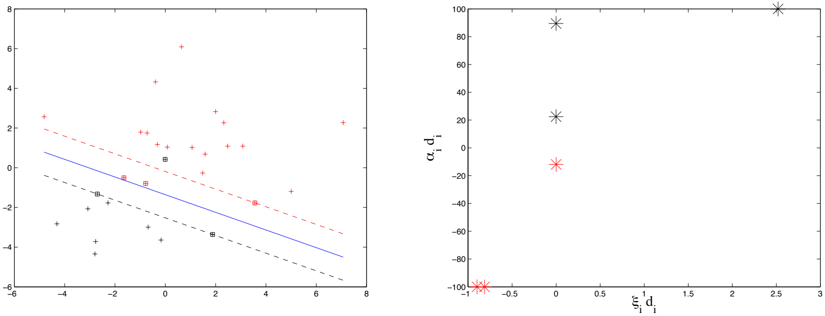


- Linearly separable data (red: negative labels).



In this case, $\alpha_n < C$.

- Nonseparable data (red: negative labels).



In this case, some α_n are equal to C.

Summary and conclusion

- The SVM is a linear machine whose criterion is to minimize the primal

$$L_p(\mathbf{w}, \xi_n) = \frac{1}{2} b||\mathbf{w}||^2 + C \sum_{n=1}^{N} \xi_n$$

subject to the constraints

$$y_n \left(\mathbf{w}^\top \mathbf{x}_n + b \right) > 1 - \xi_n, \quad \xi_n \geq 0$$

- This is equivalent to minimize the VC dimension (structural risk) and an empirical risk function.

Summary and conclusion

- The constraints define two margins

$$\mathbf{w}^\top \mathbf{x}_n + b = \pm 1$$

- The primal leads to a dual of the form

$$L_d = -\frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{YKY}\boldsymbol{\alpha}^\top + \boldsymbol{\alpha}\mathbf{1}$$

  with $\alpha_i \geq 0$, which can be optimized by quadratic programming.

- The parameter vector is expressed as a linear combination of the data

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n$$

Summary and conclusion

- The data well classified outside the margin have null values of α_n.
- The data inside the margin or misclassified have $\alpha_n = C$.
- The data on the margin have $\alpha_n < C$
- The machine is linear, though it is easy to construct nonlinear versions.
- Parameter C is free and it must be validated.
- The parameter produces a trade off between complexity and empirical risk.
- There are versions for small and medium data sizes and implementations that are capable of dealing with big data sets.